基于关键词相似度的短文本分类方法研究 *

张振豪 [1]，过 弋 [1, 2, 3†]，韩美琪 [1]，王吉祥 [1]

(1. 华东理工大学 信息科学与工程学院，上海 200237; 2. 石河子大学 信息科学与技术学院，新疆 石河子 832003; 3. 大数据流通与交易技术国家工程实验室——商业智能与可视化技术研究中心，上海 200436)

摘 要：在传统的文本分类中，文本向量空间矩阵存在“维数灾难”和极度稀疏等问题，而提取与类别最相关的关键词作为文本分类的特征有助于解决以上两个问题。针对以上结论进行研究，提出了一种基于关键词相似度的短文本分类框架。该框架首先通过大量语料训练得到 word2vec 词向量模型；然后通过 TextRank 获得每一类文本的关键词，在关键词集合中进行去重操作作为特征集合。对于任意特征，通过词向量模型计算短文本中每个词与该特征的相似度，选择最大相似度作为该特征的权重。最后选择 K 近邻(KNN)和支持向量机 SVM 作为分类器训练算法。实验基于中文新闻标题数据集，与传统的短文本分类方法相比，分类效果平均提升约 6%，从而验证了该框架的有效性。

关键词：词向量；特征选择；短文本分类；特征权重

中图分类号：TP391.1 doi: 10.19734/j.issn.1001-3695.2018.04.0440

Research on short text classification based on keyword similarity

Zhang Zhenhao[1], Guo Yi[1, 2, 3†], Han Meiqi[1], Wang Jixiang[1]

(1. *School of Information Science & Engineering, East China University of Science & Technology, Shanghai* 200237, *China*; 2. *School of Information Science & Technology, Shihezi University Xinjiang* 832003, *China*; 3. *Business Intelligence & Visualization Research Center, National Engineering Laboratory for Big Data Distribution & Exchange Technologies, Shanghai* 200436, *China*)

Abstract: In order to cope with the problem of data sparsity and “curse of dimensionality” in text classification, this paper proposes a short text classification framework by taking keyword as featrues and assigning keyword similarity as feature weight. First, it trained a word2vec model with large corpus data, then got keywords of each category text by textrank. And it selected unique keywords from the keywords collection as features. For each feature, it calculated the similarity of words in the short text by word2vec model, and assigned the maximum similarity as the weight of the feature. Finally, it chose KNN and SVM as classifier. Experiments on dataset of Chinese news headlines demonstrate that the accuracy outperforms other usual methods by 6%.

Key words: word embedding; feature selecting; short text classification; feature weighting

0 引言

近几年，由于互联网的快速发展，人们也越来越依赖于从网络中获取信息。如何快速准确地获取自己想要的信息成为当前一个重点研究课题。而文本数据量的飞速增长，混乱分布极大影响了信息获取的效率与结果，其中还包含大量诸如微博、新闻标题、商品评论等短文本。因此对短文本进行分类也吸引了越来越多的研究。

在传统的文本分类中，一般采用文本向量空间模型方法[1]，但是该方法对于短文本具有特征稀疏、语义特征不明显等特点。目前对于该问题，主要有两种方法：a)利用搜索引擎对短文本进行扩展[2]，将短文本扩展为一般文本进行分类；b)利用大量知识库、语料库作为背景知识[3]，发现词语之间的语义关系。以上两种方法均能提升段文本分类的性能。但是仍然存在着计算耗时、无法处理新词等问题。

针对特征稀疏、语义特征不明显等问题，文献[4]利用 LDA 主题—词分布矩阵的主题向量改进方法降低特征维度进行短文本分类。文献[5]融合词语类别特征和语义进行短文本分类，虽

收稿日期：2018-04-27；**修回日期：**2018-07-02 **基金项目：**国家自然科学基金资助项目（61462073）；上海市科学技术委员会项目（17DZ1101003，18511106602）

作者简介：张振豪（1993-），男），浙江杭州人，硕士研究生，主要研究方向为文本挖掘、自然语言处理；过弋（1975-），男（通信作者），教授，博士，主要研究方向为文本内容分析、知识发现与知识工程（guoyi@ecust.edu.cn）；韩美琪（1994-），女，硕士研究生，主要研究方向为文本挖掘、机器学习；王吉祥（1995-），女，硕士研究生，主要研究方向为数据挖掘、文本挖掘.

然效果相比传统方法有所提升，但是两者均采用 LDA 主题模型。该模型属于无监督学习，速度较慢，且依赖主题数量的选择，需要不断优化。文献[6]提出了提取文档关键词作为文本特征的方法，也取得了较好的分类结果。虽然文档关键词考虑了文本语义信息，但是文档关键词会随着文档数量的增加而增加，导致文本向量空间矩阵维度较大，计算耗时。本文结合不同方法的特点，采用 TextRank 提取类别关键词，作为新的短文本特征，使得文本向量空间的特征集合不再是长度达数千的词汇表而是短小有限的关键词集合。再通过 word2vec 模型计算词语之间的语义相似度作为特征权值，保留了短文本的一定语义信息，在此基础上训练分类器进行短文本分类。

本文的主要贡献如下:

a) 本文提出了一种基于类别关键词的文本特征选择方法，类别关键词个数远小于文档的总词数，能够较好地解决维度过高等问题。

b）在特征表示中本文将词语之间的相似度作为特征的权重，弥补了“词袋”模型中未考虑语义的缺点。

c) 基于以上两点本文提出了一种新的短文本分类框架，通过对真实数据集的实验表明该框架相比传统分类方法效果有一定提升。

1 相关工作与流程

设短文本集合 $\boldsymbol{D}=\{d_1,d_2,d_3,\ldots,d_n\}$，短文本类别标签 $\boldsymbol{Y}=\{y_1,y_2,y_3\ldots y_n\}$，特征集 $\boldsymbol{W}=\{w_1,w_2,w_3,\ldots,w_m\}$。本文的主要目的是通过基于关键词相似度利用 $\boldsymbol{W}$ 将 $\boldsymbol{D}$ 转换成向量空间矩阵 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\ldots,\boldsymbol{x}_n\}^T$，$\boldsymbol{x}_i(1\le i\le n)$均为 m 维向量,再训练得到分类器，能够使得将 $\boldsymbol{D}$ 中的元素尽可能分类得到正确的类别。具体流程如图 1 所示。

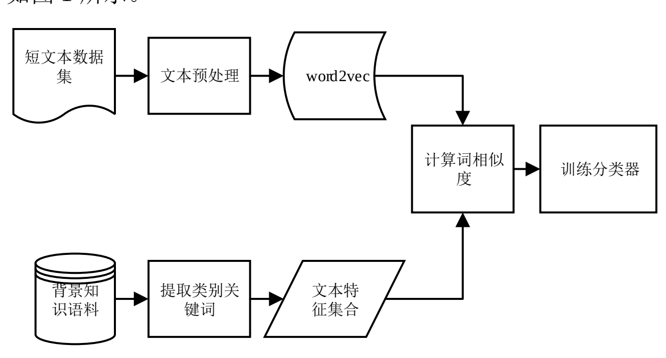


图 1 基于关键词相似度的短文本分类流程

Fig.1 Short text classification process based on keyword similarity

2 特征选择及特征表示

在本章中主要介绍基于基于 TextRank 的短文本特征选择以及基于 word2vec 的特征表示。

2.1 基于 TextRank 的短文本特征选择

传统的短文本特征的基本思想是对文本集合中的每一个特征计算某种统计度量值，并且设定一个阈值。若该度量值小于阈值，将该特征过滤；否则认为该特征为有效特征。比较典型的方法有文档频率、信息增益、互信息等。此类方法存在无法处理新文本的缺陷，即若文本中的特征均不存在有效特征集合中，该文本将无法有效表示。

针对以上问题，本文提出提取类别关键词作为文本特征。目前主流的关键词提取方法以无监督方法为主，其中涉及统计方法、图模型方法以及语义方法。以词频—逆文档频率（term frequency-inverse document frequency ,TF-IDF）[7]为基础，衍生了很多有效的关键词抽取算法，但是 TF-IDF 仅仅考虑词频，没有考虑到语义和词与词之间的关系，而且无法对单一文档进行关键词提取,并不适合进行类别的关键词提取。

TextRank[8]算法基于词汇的共现链，以图模型的方法提取文档关键词，该方法效果较好，也能够实现对单一文档的关键词提取。在 TextRank 算法中，首先构建候选关键词图 $G=(V,E,W)$，其中节点集 $V=\{v_1,v_2,v_3,\ldots,v_n\}$，由文档中的候选关键词组成，一般为词性为名词、形容词、动词的非停用词。$W=\{w_{ij}\mid 1\le i\le n\wedge 1\le j\le n\}$ 为图的权重集合，$E=\{(v_i,v_j)\mid v_i\in V\wedge v_j\in V\wedge w_{ij}\in W\wedge w_{ij}\neq 0\}$为各个节点之间的非空有限集合。两个节点之间存在边仅当它们对应的词汇在长度为 K 的窗口中共现，K 表示窗口大小，即最多共现 K 个单词。由 G 可以得到对应的相似度矩阵 $S_{n\times n}$，如式(1)所示。

$$S_{n\times n}=\begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{pmatrix} \tag{1}$$

由 G 和 $S_{n\times n}$，通过式(2)迭代计算各节点的权重。

$$WS(v_i)=(1-d)+d\times\sum_{v_j\in In(v_i)}\frac{w_{ji}}{\sum_{v_k\in Out(v_j)}w_{jk}}WS(v_j) \tag{2}$$

其中：$WS(v_i)$为节点 vi 的权重值；d 是阻尼系数，一般设置为 0.85，表示任意一节点跳转到其他任意节点的概率；$In(v_i)$表示指向节点 v_i 的所有节点的集合；$Out(v_i)$表示 v_i 指向的所有节点的集合。一般当迭代 20~30 次，迭代阈值设置为 0.000 1。计算结束之后，根据节点的权重降序排列。在此基础上，本文根据每类文本的背景知识语料提取类别关键词，如对于体育类背景知识语料，提取的关键词依次为{“比赛”，“球队”，“球员”…..}等、对于财经类背景知识语料、提取的关键词依次为{“中国”，“公司”，“市场”…..}等。

2.2 基于 word2vec 的特征权重

在确定文本特征之后，需要对每一个样本的所有特征分配一定的权重。分配权重的好坏将极大影响分类效果。比较传统的权值分类方法有二分类（该特征在文本出现即为 1，未出现即为 0）、词频(tf)、逆文档频率(idf)、TF-IDF 等。文献[9]通过文档中不包含的词计算存在的词的特征权重，在其他条件相同的情况下，较之前的 TF-IDF，信息熵等方法分类效果有明显的提升。文献 [10]针对朴素贝叶斯分类器，通过从训练数据深度计算特征加权频率一定程度提高了分类效果。然而大多数有关特征权重的研究还是仅仅考虑特征频率等“词袋”层面的联系，

未考虑到特征与文本之间的语义关系。为此，本文将词与词之间的相似度作为特征权值，在文本转换为向量的过程中保留了一定语义内容。

为了准确计算词语之间的相似度，本文选择 word2vec[11] 模型。Word2vec 模型自提出以来就得到了广泛应用，文献[12]利用 word2vec 模型用于观点分类，效果比词袋模型出色。文献[13]借助 word2vec 模型所包含的语义信息提取中文评论的情感特征，在情感分析方面取得了不错的效果。Word2vec 模型其本质是一种具有隐含层的神经网络，输入输出均为词汇表，通过学习词与上下文之间的关系，待神经网络收敛之后，从输入层到隐含层的向量代表词汇表中每个词的向量。Word2vec 包含两种训练模型 CBOW(continueous bag-of-words)和 Skip-Gram。相对来说 CBOW 更适用于较大的语料数据。因此本文采用 CBOW 训练模型得到的词向量模型表示短文本中的词汇。

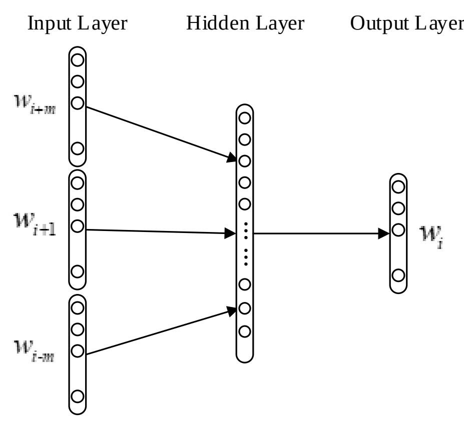


图 2　CBOW 模型

Fig.2　CBOW model

在 CBOW 训练模型中，它将词的上下文作为输入，而该词作为输出，如图 2 所示。输入层为第 i 个词的前后 m 个词，输出层为第 i 个词 w_i，输入输出层向量均为 one-hot encoding，即维度数等于词汇数量。核心思想即根据上下文预测当前词语的概率。在训练结束以后，可以得到所有词的词向量。令第 i 个词的词向量为 s_i，第 j 个词的词向量为 s_j，则可以根据余弦相似度计算得到两个词之间的相似度，如式(3)所示。

$$sim(s_i,s_j)=\frac{s_i \bullet s_j}{|s_i||s_j|} \tag{3}$$

3　基于关键词相似度的短文本分类方法

本章将介绍基于关键词相似度的短文本分类方法的具体框架以及相关细节。各符号定义如表 1 所示。

表 1　短文本分类中各符号的定义

Table 1　Definition of symbols

符号	定义
$P=\{p_1,p_2,\ldots,p_k\}$	不同类别的背景知识语料
$D=\{d_1,d_2,\ldots,d_n\}$	短文本集合
C	短文本类别集合
$d_i=\{q_1,q_2,\ldots,q_m\}$	文本 i 的词汇集合
$K=\{k_1,k_2,\ldots,k_k\}$	所有类别的关键词集合
k_j	类别 j 的关键词集合
Y	短文本的标签集合
Dic	文本特征集合
$V_i=\{w_1,w_2,..,w_q\}$	文本 i 表示的特征向量
sim(a,b)	向量 a 与向量 b 的相似度
Model	词向量模型

3.1　短文本特征选择

基于 TextRank，本文依据算法 1 获得短文本的文本特征。

算法 1　基于关键词的短文本特征选择

输入：不同类别的文本语料 P，关键词个数 n。

输出：类别关键词集合 Dic。

① for p_i in P:

② 　p_i=Text pre-process(p_i);/*对背景知识语料做文本预处理*/

③ 　k_i=TextRank(p_i,topk=n)/*TextRank 降序排列关键词，选取第 i 类语料的前 n 个关键词*/

④ end for

⑤ for k_i in K:

⑥ 　for each word in k_i:

⑦ 　　if word appears only once:

⑧ 　　　Dic.append(word)/*对各类别关键词集合去重，保留仅在一个类别关键词集合中出现的词*/

⑨ 　end for

⑩ end for

3.2　短文本特征向量表示

在进行文本分类之前，都需要将文本转换成特征向量以便能够使用分类器进行训练和学习。本文的特征向量表示方法如算法 2 所示。

算法 2　短文本集合特征向量表示

输入：文本特征集合 Dic，词向量模型 Model，短文本集合 D。

输出：短文本集合特征向量 V。

① for d_i in D:

② 　for each w in Dic:/*Dic 中元素个数即为特征个数*/

③ 　　max_sim =-1

④ 　　for q_j in d_i:

⑤ 　　　s=sim(Model(q_j),Model(w))/*通过词向量将词转换成向量进行相似度计算*/

⑥ 　　　if(s>max_sim)max_sim =s

⑦ 　　end for

⑧ 　w_j=max_sim/*文本 i 的第 j 个特征权重*/

⑨ 　end for

⑩ end for

3.3 短文本分类

通过以上方法就能够得到短文本集合的特征向量矩阵 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\ldots,\boldsymbol{x}_n\}^T$，$n$ 为短文本数据集大小；$\boldsymbol{x}_i(1\le i\le n)$为 m 维向量；m 为特征集大小。结合标签集合 Y 可以使用分类算法如 KNN、SVM 等进行实验，验证以上特征选择和特征权重计算方法的优劣。

4 实验与结果

为了验证基于关键词相似度的短文本分类方法的有效性，本文利用近两年的新浪热点新闻数据进行实验。首先介绍数据集以及对比的算法；最后在数据集上进行 10 折交叉验证该方法的有效性。

4.1 数据集与实验设置

本文以近两年的热点新闻为实验数据，采集了包括体育、社会、娱乐、财经、科技、国际、军事七大类的新闻标题以及正文。标题用于短文本分类，正文用于作为背景知识进行关键词提取。同时为了数据均匀分布，各类新闻标题数据数量均为 2 000 条，如表 2 所示。而 word2vec 模型则由额外语料（百度百科 20 G、新闻语料 12 G、小说 90 G）数据训练而得。对于分类算法的设置中，SVM 核函数采用径向基函数(radial basis function，RBF)。RBF 在线性不可分的情况下效果优于线性核且计算耗时少于多项式核。通过实验比较，KNN 选择近邻数 N 为 15 较合适。

表 2　新闻标题数据集

Table 2　Data set of news headline

Category	Number of news headlines
Technology	2000
Sports	2000
Society	2000
Entertainment	2000
Military	2000
International	2000
Finance	2000

4.2 实验设计

为验证本文提出的方法的有效性，首先在相同的分类算法上，验证不同的关键词个数对分类效果的影响；然后选择较合适的关键词个数作为实验参数，与以下两个实验做对比。

a)TF-IDF　对数据进行常规预处理（去除标点符号、去除停用词等）后，计算短文本的 TF-IDF 特征向量。最后使用 KNN 和 SVM 进行训练和验证。

b) Sum-CBOW　由于短文本词数较少，获得的 TF-IDF 特征向量维度高且极其稀疏，且词向量本身具有较好的语义特征，文献[14]提出了将短文本中所有词的词向量累加作为短文本的特征向量的方法，即特征维度为词向量的维度而不是类别关键词的数目。该方法的文本向量空间维度低、分类速度快，效果也相对较好。该方法在本文中简称 Sum-CBOW。

最后为验证语义性对分类效果的提升，在特征集保持不变的情况下，特征权重由词频代替，对比实验结果。在此过程中，本文提出的方法为 Key-CBOW。

4.3 实验评估

本次实验评估分别有以下四个指标:

a) 分类准确率 Precision。

类别 c_i 的分类准确率 p_i，如式(4)所示。

$$p_i=\frac{\text{分类结果中正确分为}c_i\text{的样本个数}}{\text{分类结果中所有分为}c_i\text{的样本个数}} \tag{4}$$

b) 分类召回率 Recall。

类别 ci 的分类召回率 ri，如式(5)所示。

$$r_i=\frac{\text{分类结果中正确分为}c_i\text{的样本个数}}{\text{类别为}c_i\text{的实际样本个数}} \tag{5}$$

c) F1 分数。

类别 ci 的 F1 分数 f1i，如式(6)所示。

$$f1_i=\frac{2\times p_i\times r_i}{p_i+r_i} \tag{6}$$

d）宏平均 F1 分数。

它是所有类别的 F1 分数的算术平均值，如式(7)所示。

$$F1_{macro}=\frac{1}{k}\sum_{i=1}^{k}f1_i \tag{7}$$

4.4 实验结果及分析

图 3 验证了类别关键词个数对短文本分类效果的影响，验证指标为 F1-macro。关键词个数取 N={ 20, 40, 60, 80, 100, 120, 140, 160}。总体上，SVM 算法要优于 KNN 算法，原因在于特征为类别关键词，类别间可分性比较好；而 KNN 比较适合基于样本相似度的方法，KNN 并不依赖于特征的可区分度。当关键词个数较少时，效果均不是很理想，随着关键词个数的增加，文本特征集扩大，分类效果提升。当关键词个数达到一定值之后达到稳定，F1-macro 值不再提高。为避免过拟合以及欠拟合等问题，本文后续实验中的关键词个数均设置为 100。

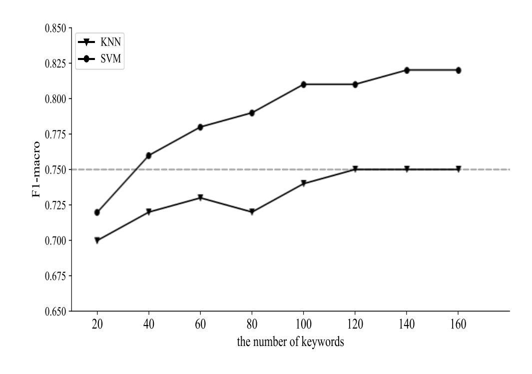


图 3　基于关键词个数的 F1-macro 值对比

Fig.3　Comparison of F1-macro based on number of keywords

对比实验结果如表 3 所示。根据表 3 的结果可以看到，Sum-CBOW 的分类效果相对较差，将短文本的词向量直接叠加作为该文本的特征向量，虽然能够一定程度保留该短文本的语义信息，但削弱了类别特征的区分度。传统的 TF-IDF 在本次

实验中效果较好，但是存在召回率和准确率偏差略大、特征多、维度稀疏、计算耗时等问题。而本文方法选取类别关键词作为特征以及将短文本与特征的最大相似度作为特征权重，能够有效挖掘出短文本与不同类别的核心语义关联，在转换为特征向量的过程中保留语义。因此对于测试数据集，虽然部分指标未达到最大值，但是本文方法在准确率和召回率保持平衡的同时，将平均准确率和平均召回率都提升了约 6%。图 4 为三个实验对于数据集各类别分类 F1 值的直观对比，证明了本文提出的基于关键词相似度的短文本分类方法能够有效提高分类效果。

表 3　不同算法的分类效果表现比较

Table 3　Comparison of classification effects on different algorithms

	TF-IDF			Sum-CBOW			Key-CBOW		
Category	P	R	F1	P	R	F1	P	R	F1
Technology	0.77	0.74	0.76	0.58	0.75	0.66	0.74	0.73	0.73
Sports	0.92	0.88	0.90	0.91	0.87	0.89	**0.94**	**0.95**	**0.95**
Society	0.57	0.71	0.64	0.80	0.51	0.63	0.75	0.71	**0.73**
Entertainment	0.70	0.90	0.79	0.68	0.91	0.78	**0.89**	0.90	**0.90**
Military	0.80	0.87	0.83	0.79	0.81	0.80	**0.83**	0.86	**0.85**
International	0.75	0.63	0.69	0.68	0.91	0.78	0.73	0.77	0.75
Finance	0.88	0.55	0.68	0.82	0.52	0.64	0.76	**0.74**	**0.75**
Average	0.77	0.75	0.75	0.75	0.75	0.74	**0.81**	**0.81**	**0.81**

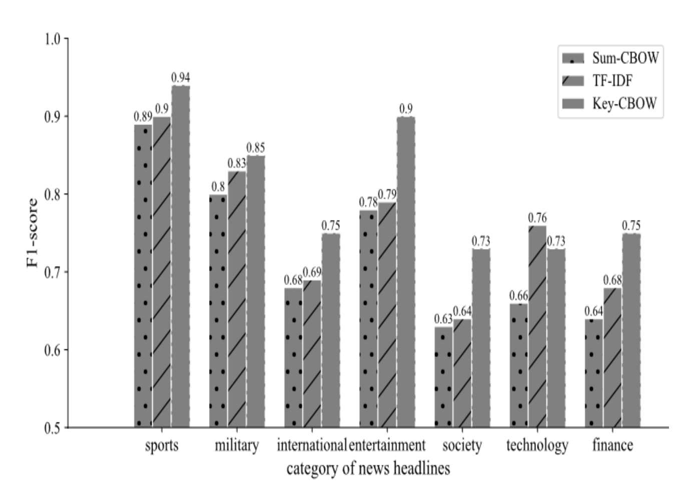


图 4　不同算法的类别 F1 值对比

Fig.4　Comparison of F1 values of category of different algorithms

针对验证语义关系对分类能力的影响，本文设计了第三个实验 Key-Frequency。在此实验中，对于短文本特征向量 $x=\{w_1,w_2,\ldots,w_m\}$（m 为特征集个数），特征权重不再是短文本与该特征的相似度，而是该特征在短文本中的频次，KNN 为该实验的分类算法。实验结果的 F1 值如表 4 所示。TF-IDF 分类效果明显比该方法出色，说明在缺失了语义相似度和特征较少的情况下，以词频为特征权重无法处理大部分且不在特征集合中的词，从而验证了本文提出的将语义相似度作为特征权重的方法能够有效提升分类效果。

表 4　语义相似度对分类 F1 值的影响

Table 4　Influence of semantic similarity on F1 value

Category	Key-Frequency	TF-IDF
Technology	0.31	0.76
Sports	0.40	0.90
Society	0.44	0.64
Entertainment	0.33	0.79
Military	0.55	0.83
International	0.10	0.69
Finance	0.31	0.68

5　结束语

鉴于短文本分类中高度稀疏、缺少语义特征等问题，本文提出了一种基于关键词相似度的短文本分类框架。该框架综合考虑类别词汇以及语义信息，通过选择有限的类别关键词作为特征集合以及将词与特征之间的相似度作为特征权重，既解决了文本特征维度过高的问题，又保留了短文本的语义，提高了文本的区分度。本文通过基于词频和基于语义的对比实验，验证了该框架的有效性。

接下来的工作主要是分析不同类别短文本分类效果差异较大的现象，以及考虑相应的优化方法。对于用语义相似度表示特征权重的方法，可以考虑对比语义相似度的不同计算方法，验证本文方法的可行性。

参考文献：

[1] Salton G, Wong A, Yang C S. A vector space model for automatic indexing [M]. New York: ACM Press, 1974.

[2] Sun Aixin. Short text classification using very few words [C]// Proc of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2012: 1145-1146.

[3] Chen Mengen, Jin Xiaoming, Shen Dou. Proceedings of the twenty-second international joint conference on artificial intelligence short text classification improved by learning multi-granularity topics [J]. Journal of Shanxi Institute of Economic Management, 2012: 152-157.

[4] 杨萌萌，黄浩，程露红，等. 基于 LDA 主题模型的短文本分类 [J]. 计算机工程与设计，2016, 37 (12): 3371-3377. (Yang Mengmeng, Huang Hao, Cheng Luhong, *et al.* Short text classification based on LDA topic model [J]. Computer Engineering and Design, 2016, 37 (12): 3371-3377.)

[5] 马慧芳，周汝南，吉余岗，等. 融合词语类别特征和语义的短文本分类方法 [J]. 计算机工程与科学，2017, 39 (2): 399-404. (Ma Huifang, Zhou Runan, Ji Yugang, *et al.* A short text classification method combining lexical category features and semantics [J]. Computer Engineering & Science, 2017, 39 (2): 399-404.)

[6] Onan A, Korukoğlu S, Bulut H. Ensemble of keyword extraction methods and classifiers in text classification [J]. Expert Systems with Applications, 2016, 57 (C): 232-247.



[7] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval [M]. Oxford: Pergamon Press, Inc, 1988.

[8] Mihalcea R. TextRank: bringing order into texts [J]. Empirical Methods on Natural Language Processing, 2004: 404-411.

[9] Sabbah T, Selamat A, Selamat M H, *et al.* Modified frequency-based term weighting schemes for text classification [J]. Applied Soft Computing, 2017, 58: 193-206.

[10] Jiang Liangxiao, Li Chaoqun, Wang Shasha, *et al.* Deep feature weighting for naive Bayes and its application to text classification [J]. Engineering Applications of Artificial Intelligence, 2016, 52 (C): 26-39.

[11] Mikolov T, Sutskever I, Chen Kai, *et al.* Distributed representations of words and phrases and their compositionality [J]. 2013, 26: 3111-3119.

[12] Enríquez F, Troyano J A, López-Solaz T. An approach to the use of word embeddings in an opinion classification task [J]. Expert Systems with Applications, 2016, 66: 1-6.

[13] Zhang Dongwen, Xu Hua, Su Zengcai, *et al.* Chinese comments sentiment classification based on word2vec and SVM perf [J]. Expert Systems with Applications, 2015, 42 (4): 1857-1863.

[14] 江大鹏. 基于词向量的短文本分类方法研究 [D]. 杭州: 浙江大学, 2015. (Jiang Dapeng, Research on short text classification based on word distributed representation [D]. Hangzhou: Zhejiang University, 2015.)